CLAIMS

1. A method for extending sparse alignment coverage for any combination of RNA derived sequence fragments, comprising the steps of:

combining and catenating all combinations of overlapping alignments that agree with each other; and

extending boundaries of overlapping alignments that agree with their first and last exons.

2. The method of Claim 1, further comprising the step of:

preprocessing and filtering.

3. The method of Claim 2, wherein said preprocessing and filtering step comprises the steps of:

reading a file containing alignments of exon annotations to genomic sequence; and

populating an array of data structures, one for each alignment, wherein each alignment is stored as a set of alignment blocks, each block representing a matching region between a given RNA and the genomic.

4. The method of Claim 3, wherein said blocks are considered exons, and gaps between them are considered introns.

5. The method of Claim 4, wherein 5'/3' splice sites are referred to as hard edges, and the two ends of an alignment are referred to as soft ends.

6. The method of Claim 3, said preprocessing and filtering step further comprising the steps of:

determining if gaps are introns or inserts;

eliminating single exon alignments;

trimming soft ends of alignments; and

filtering out highly similar alignments.

7. The method of Claim 5, wherein gaps in alignment blocks that are less than or equal to twenty nucleotides are considered inserts instead of introns, and wherein two adjacent blocks are subsequently combined into one.

8. The method of Claim 7, wherein once an entire alignment has been read and all inserts removed, an alignment is discarded if it consists of only a single exon.

9. The method of Claim 8, wherein soft ends of an alignment are trimmed by ten nucleotides.

10. The method of Claim 9, wherein once all multi-exon alignments have been read and trimmed, there is a filtering step to cut down on running time.

11. The method of Claim 1, further comprising the step of:

discarding similar alignments, wherein two alignments are similar if they are on a same strand and have similar number of agreeing exons.

12. The method of Claim 1, further comprising the step of:

eliminating shorter alignments.

13. The method of Claim 5, further comprising the step of:

removing similar alignments only if their soft ends are not on opposing sides of some other alignment's hard edge.

14. The method of Claim 11, further comprising the step of:

whenever a similar alignment is eliminated, an alignment that remains is stretched to widest possible soft end points of said two alignments to keep said alignments as long as possible, wherein for a set of many similar alignments whose soft ends are not near any hard edges, only one alignment remains after said filtering step; and wherein soft ends of a remaining alignment are widest possible among all of similar alignments.

15. The method of Claim 1, wherein said RNA derived sequence fragments comprise any of EST's, partial cDNA's and full-length cDNA's.

16. A method for extending sparse alignment coverage for any combination of RNA derived sequence fragments, comprising the steps of:

merging all combinations of overlapping alignments that agree with each other; and

extending boundaries of overlapping alignments that agree with their first and last exons.

17. The method of Claim 16, wherein said merge step performs a pairwise comparison of each overlapping RNA sequence on a same strand;

wherein if their splice sites agree, then either a new, longer alignment is created out of said two sequences, or one of said alignments is labeled as redundant to the other and it is scheduled for deletion.

18. The method of Claim 17, wherein an alignment X is redundant if it agrees with another alignment Y, and soft ends of X fall completely within soft ends of Y, inclusive.

19. The method of Claim 18, wherein redundant alignments are not deleted until after said pairwise comparisons are complete.

20. The method of Claim 19, wherein a smaller alignment can be redundant with respect to another larger alignment and still be able to merge with other alignments that a larger alignment cannot; and wherein redundant pieces are still available for said pairwise comparisons, even though they are redundant with other alignments.

21. The method of Claim 20, wherein each newly created, merged alignment is also checked against other, unmerged alignments in a same way, either merging once again, becoming labeled as redundant, or causing other alignments to be labeled redundant.

22. The method of Claim 21, wherein whenever an alignment, merged or not, has been compared to all other alignments, and is neither redundant nor merged it is placed on a list of Done alignments.

23. The method of Claim 22, wherein once all comparisons have been finished, said Done alignments are again compared pairwise to check for redundancies.

24. The method of Claim 16, wherein said merging step comprises the following algorithm:

```
/       * cdnaAray = array of cDNA alignments stored by increasing
        *                   start points
        * todo list = a list of merged alignments that are
        *                   scheduled to be compared to the other          *
                    alignments in cdnaArray. Associated with each
        *                   merged alignment is an integer
        *                   array index indicating where in cdnaArray the
        *                   comparisons should begin.
    */
bool merged;
for I = 0 to sizeOf (cdnaArray) {
        for j = (I + 1) to sizeOf (cdnaArray) {
                merged = 0;
        if  ( overlap (cdnaArray[I], cdnaArray[j])) {
                        merged I = merge (cdnaArray[I], cdnaArray[j], todo_list);
        }
      }
      if (!merged) {
         push(Done_list, cdnaArray[I]);
      }
      else {/* there must be a mered alignment on the todo_list */
        while (todo_list !=NULL) {
                todo = pop(todo_list);
                merged = 0;
                for k = todo,dtart to size Of (cdnaArray) {
                if (overlap (todo, cdnaArray [k])) {
                        merged I=merge (todo, cdnaArray [k], todo_list);
                }
```

```
                }
                if (!merged) {
                push(Done-list, todo);
            }
                }
        }
}.
```

25. The method of Claim 16, wherein said extending step comprises the step of:

performing a pairwise comparison of all remaining alignments, wherein for each pair, left-most and right-most overlapping exons are considered, and wherein if there are no conflicts on any of their hard edges, then short ended alignments are extended in such a way that they match longer alignments.

26. The method of Claim 25, wherein if an alignment can be extended, then a new alignment is created, and it is further compared against all overlapping alignments, and wherein if an alignment cannot be extended, it gets placed on a Done list.

27. The method of Claim 26, wherein after all comparisons are complete, said Done list is again purged of redundant alignments.

28. The method of Claim 16, wherein said RNA derived sequence fragments comprise any of EST's, partial cDNA's and full-length cDNA's.

29. An apparatus for extending sparse alignment coverage for any

combination of RNA derived sequence fragments, comprising:

a computer implemented algorithm for combining and catenating all combinations of overlapping alignments that agree with each other; and

a computer implemented algorithm for extending boundaries of overlapping alignments that agree with their first and last exons.

30. The apparatus of Claim 29, further comprising:

computer implemented means for preprocessing and filtering.

31. The apparatus of Claim 30, wherein said preprocessing and filtering means comprises means for:

reading a file containing alignments of exon annotations to genomic sequence; and

populating an array of data structures, one for each alignment, wherein each alignment is stored as a set of alignment blocks, each block representing a matching region between a given RNA and the genomic.

32. The apparatus of Claim 31, wherein said blocks are considered exons, and gaps between them are considered introns.

33. The apparatus of Claim 32, wherein 5'/3' splice sites are referred to as hard edges, and the two ends of an alignment are referred to as soft ends.

34. The apparatus of Claim 33, said preprocessing and filtering means further comprising means for:

determining if gaps are introns or inserts;

eliminating single exon alignments;

trimming soft ends of alignments; and

filtering out highly similar alignments.

35. The apparatus of Claim 34, wherein gaps in alignment blocks that are less than or equal to twenty nucleotides are considered inserts instead of introns, and wherein two adjacent blocks are subsequently combined into one.

36. The apparatus of Claim 35, wherein once an entire alignment has been read and all inserts removed, an alignment is discarded if it consists of only a single exon.

37. The apparatus of Claim 36, wherein soft ends of an alignment are trimmed by ten nucleotides.

38. The apparatus of Claim 37, wherein once all multi-exon alignments have been read and trimmed, there is a filtering step to cut down on running time.

39. The apparatus of Claim 29, further comprising:

means for discarding similar alignments, wherein two alignments are similar if they are on a same strand and have similar number of agreeing exons.

40. The apparatus of Claim 29, further comprising:

means for eliminating shorter alignments.

41. The apparatus of Claim 33, further comprising:

means for removing similar alignments only if their soft ends are not on opposing sides of some other alignment's hard edge.

42. The apparatus of Claim 39, wherein whenever a similar alignment is eliminated, an alignment that remains is stretched to widest possible soft end points of said two alignments to keep said alignments as long as possible, wherein for a set of many similar alignments whose soft ends are not near any hard edges, only one alignment remains after said filtering step; and wherein soft ends of a remaining alignment are widest possible among all of similar alignments.

43. The apparatus of Claim 29, wherein said RNA derived sequence fragments comprise any of EST's, partial cDNA's and full-length cDNA's.

44. An apparatus for extending sparse alignment coverage for any combination of RNA derived sequence fragments, comprising:

a computer implemented algorithm for merging all combinations of overlapping alignments that agree with each other; and

a computer implemented algorithm for extending boundaries of overlapping alignments that agree with their first and last exons.

45. The apparatus of Claim 44, wherein said merge algorithm performs a pairwise comparison of each overlapping RNA sequence on a same strand; wherein if their splice sites agree, then either a new, longer alignment is created out of said two sequences, or one of said alignments is labeled

as redundant to the other and it is scheduled for deletion.

46. The apparatus of Claim 45, wherein an alignment X is redundant if it agrees with another alignment Y, and soft ends of X fall completely within soft ends of Y, inclusive.

47. The apparatus of Claim 46, wherein redundant alignments are not deleted until after said pairwise comparisons are complete.

48. The apparatus of Claim 47, wherein a smaller alignment can be redundant with respect to another larger alignment and still be able to merge with other alignments that a larger alignment cannot; and wherein redundant pieces are still available for said pairwise comparisons, even though they are redundant with other alignments.

49. The apparatus of Claim 48, wherein each newly created, merged alignment is also checked against other, unmerged alignments in a same way, either merging once again, becoming labeled as redundant, or causing other alignments to be labeled redundant.

50. The apparatus of Claim 49, wherein whenever an alignment, merged or not, has been compared to all other alignments, and is neither redundant nor merged it is placed on a list of Done alignments.

51. The apparatus of Claim 50, wherein once all comparisons have been finished, said Done alignments are again compared pairwise to check for

redundancies.

52. The apparatus of Claim 44, wherein said merging algorithm comprises the following:

```
/       * cdnaAray = array of cDNA alignments sorted by increasing
        *            start points
        * todo list = a list of merged alignments that are
        *            scheduled to be compared to the other        *
              alignments in cdnaArray. Associated with each
        *            merged alignment is an integer
        *            array index indicating where in cdnaArray the
        *            comparisons should begin.
    */
bool merged;
for I = 0 to sizeOf (cdnaArray) {
        for j = (I + 1) to sizeOf (cdnaArray) {
              merged = 0;
        if  ( overlap (cdnaArray[I], cdnaArray[j])) {
              merged I = merge (cdnaArray[I], cdnaArray[j], todo_list);
        }
      }
      if (!merged) {
        push(Done_list, cdnaArray[I]);
      }
      else {/* there must be a mered alignment on the todo_list */
        while (todo_list !=NULL) {
              todo = pop(todo_list);
              merged = 0;
              for k = todo,dtart to size Of (cdnaArray) {
              if (overlap (todo, cdnaArray [k])) {
                    merged I=merge (todo, cdnaArray [k], todo_list);
              }
              }
              if (!merged) {
              push(Done-list, todo);
```

```
        }
    }
}
}.
```

53. The apparatus of Claim 44, wherein said extending step comprises the step of:

performing a pairwise comparison of all remaining alignments, wherein for each pair, left-most and right-most overlapping exons are considered, and wherein if there are no conflicts on any of their hard edges, then short ended alignments are extended in such a way that they match longer alignments.

54. The apparatus of Claim 53, wherein if an alignment can be extended, then a new alignment is created, and it is further compared against all overlapping alignments, and wherein if an alignment cannot be extended, it gets placed on a Done list.

55. The apparatus of Claim 54, wherein after all comparisons are complete, said Done list is again purged of redundant alignments.

56. The apparatus of Claim 44, wherein said RNA derived sequence fragments comprise any of EST's, partial cDNA's and full-length cDNA's.